DOCUMENT RESUME

ED 122 603 FL 007 521

AUTHOR	Moller, Alan
TITLE	Validity in Proficiency Testing. ELT Documents (75/3).
INSTITUTION	British Council, London (England). English-Teaching Information Centre.
PUB DATE	27 Jun 75
NOTE	17p.
AVAILABLE FROM	English Teaching Information Centre, The British Council, 10 Spring Gardens, London SW1A 2BN, England
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS	Adult Students; *English (Second Language); Foreign Students; *Language Proficiency; Language Skills; *Language Tests; Measurement Instruments; Performance Tests; Test Construction; *Test Validity
IDENTIFIERS	*Davies Test

ABSTRACT

Some validity criteria were applied to the language proficiency measures used by the British Council, particularly the "Davies Test." Five types of test validity are identified: construct validity; content validity; concurrent validity compared with another test; prediction validity; and face, or apparent, validity. Foreign students studying in Britain were tested for proficiency in English for academic and personal uses. In addition, students' tutors were asked for assessments, and the tester evaluated students' oral and written communicative skills and comprehension. Correlations were made between students' performance assessment prior to departure for Britain and tutors' ratings after arrival, and the first assessment was found about 85 percent accurate. It was concluded that the Davies Test, despite weaknesses of construct, is essentially a valid instrument. Subjective assessments, especially when rigorously administered, were generally valid. Students all improved when in Britain and most passed their threshold of minimum necessary proficiency. It is proposed that a carefully constructed and tested proficiency measure is likely to possess overall validity even if some parts are deficient. (CHK)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED122603

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

VALIDITY IN PROFICIENCY TESTING - Alan Moller.
(This paper summarises certain points and expands others presented by the writer in a seminar on this topic held in the English-Teaching Information Centre on 27 June 1975)

1. Introduction

1.1 The primary concern of any test constructor is that his test should be a valid one. He cannot sit back and relax until he is satisfied that there are clear indications that the instrument recently developed possesses validity. This is particularly true of the writer of a language proficiency test, for in his search for validity he will have to define, or at least imply what is meant by, proficiency in a particular language. It is likely that he will settle for a definition at a very general level - and thereby make it more difficult to determine the extent of the validity of his measures.

1.2 Validity, however, is not solely the concern of the test constructor. It is very much the concern of the persons being tested and of the organisations who require information on the candidates tested. When a test has been developed, details of the procedures followed to establish its validity are usually published for the information of the test users. But, whatever the claims of the test constructors, many of the test users will query its validity after they have used it a certain number of times. In some cases this may be because the test is inappropriate to their needs or merely the next best instrument available, but in other cases users may well feel that the test is not yielding the information required. In such a case a close study of the test performance is called for.

1.3 In considering this topic, it would be instructive to look briefly at the following questions:

What is proficiency?
Is proficiency testing different from other language testing?
What is validity?

FL007521

and then apply some validity criteria to the proficiency measures used by the British Council, particularly the 'Davies Test', and reach some conclusions.

2. Language proficiency

Proficiency is essentially concerned with an individual's ability to use a language appropriately and effectively. It is the use of language in 'real life situations', as Clark (1972) terms it. It is not merely displaying one's knowledge of a language but displaying one's ability to use it in a variety of situations. If this is accepted, then it is important to know what those situations are, or what they are likely to be. In other words, proficiency for what? In the context of this discussion, therefore, the situations in which the student is required to be proficient are broadly those in which English is used as a medium for further tertiary level education in Britain - English for Academic Purposes (EAP). In addition the student will need to use English for the purpose of everyday living in Britain, the environment in which he will be pursuing his studies. This I term English for personal purposes (EPP). Proficiency in English therefore, might be characterised as the ability to express oneself, to understand discourse, and to participate or interact effectively in real life situations through the medium of English.

3. Proficiency testing

3.1 Proficiency testing is a form of attainment testing together with achievement and diagnostic testing. It is worth maintaining the distinctions between the different types of attainment tests since each reflects a different purpose. Achievement tests are based on a particular course of study or syllabus and are devised to find out how much of the language material that has been presented has in fact been mastered by the students. Diagnostic testing is 'in-depth' testing, and the test constructors will ensure that sufficient items are given on the selected features of the language to be tested so that the students' strengths and weaknesses can be ascertained. Because of constraints of time and manpower a proficiency test is likely to be shorter. It will contain items that discriminate well, and it will not be based on any particular course but on the test constructors' decisions as to the appropriate language to be sampled and as to the standard which can be reasonably expected of the candidates. A proficiency test is an external test, one which is imposed on a candidate from outside and which does not necessarily arise from his immediate studies or daily occupation. It should relate to some future need to use the language in certain situations.

3.2 The question now arises as to the type of information on the candidate's proficiency that is required. Do we wish to compare his performance with that of other candidates or with a particular group - presumably people whose language needs are or were similar? If this is so, a norm-referenced proficiency test is required, and we shall need to know or determine the characteristics of the norm group. Or do we wish to match the candidate's performance with a set of criteria? Has he or she reached a certain standard? In this case we shall need to know or determine the criteria, and a criterion-referenced measure will be required. A single norm-referenced test may be suitable as a criterion-referenced test, but it is important to be clear as to what type of information the test is to yield.

4. Validity

4.1 There is a range of literature in the field of psychology and measurement in education dealing with the concept of validity, and

Pilliner's contribution in the 'Language Testing Symposium' (ed. Davies, 1968) is very helpful. In essence when we seek to establish the validity of a test we are attempting to find out the extent to which the test is measuring what it is presumed to measure, how well it is doing the job that is required of it; how much of the required information it is yielding.

4.2 Five types of validity are commonly identified:

4.2.1 Construct validity. A test may be deemed to possess construct validity if all the factors that make up the variable or behaviour being tested are present. In language proficiency testing we are faced with determining the factors that go to make up that proficiency. Can we isolate the factors? How discrete are they? How far are they combined or integrated? Carroll (1961) suggested a matrix of language competencies with listening, speaking, reading, writing on the one axis and phonology, lexicon, grammar on the other. As the various subtests are constructed they could be checked off in the matrix eg listening/phonology, reading/grammar, reading/lexicon. Similarly in constructing, say, a listening comprehension test we should need to determine what the factors are and incorporate them into our test or series of tests. Would it include segmental listening, phoneme discrimination? Would it include short utterances, dialogues, lengthy monologues, and so on?

4.2.2 Content validity relates to the appropriacy of the language sampled by the test. In an achievement test, based on a specific course of study or set of materials, the language to be tested is clear, since it is contained in the course outline. But in the case of a proficiency test the guidelines are not so clear. The general framework of English for Academic Purposes is still extremely broad and the language functions varied. The test constructors have to decide which types of discourse to include, which features of grammar, which areas of lexis, and the contexts in which they are to be set. The potential test consumer must decide whether the final selection is acceptable. The language presented in the course of one or one and a half hours of testing is going to be but a minute sample.

4.2.3 Concurrent validity. Once the construct and the content have been decided and the test is ready for piloting it is advisable to check the results of the new test against results yielded by another, external criterion measure. In language proficiency testing this may take the form of another, hopefully valid, proficiency test. More often it will be in the form of a rating of the candidates by a teacher or other person who knows their abilities well. The teacher may rank a group in order of ability based on his subjective judgment, or he may base his ranking on marks obtained by the members of the group over a period of instruction. Although there may be doubts about the validity or reliability of some of these measures, if given at the same time as the piloting of the test being validated, they should provide enough information to reach some conclusion on concurrent validity.

4.2.4 Predictive validity. This type of validity calls for another, second, external criterion which can only be applied after some time has elapsed following the administration of the

test. It will involve an assessment of the subsequent performance of candidates in the activity or situation in which they engage after taking the language test. This is particularly important in the validation of a language proficiency test, although performance in any language test is bound to imply a prediction of future performance by the candidate. Thus if the purpose of a test is to predict future academic success, or lack of it, or to predict future language ability under certain conditions, evidence of subsequent academic or linguistic achievement is necessary in order to have some indication of a test's predictive power.

4.2.5 Face validity. This is an aspect of validity which is particularly important for candidates, and their teachers. Does the test appear to be valid? Do the tasks they are required to perform make sense in a language proficiency test? For example, if studies were to show that language ability and length of hair grown over a period correlated very highly, it might not be unreasonable to require candidates to submit to a procedure of measurement of hair growth! Such a procedure would hardly seem relevant and so would lack face validity! On the other hand a candidate is quite prepared for conversations and sounds to be played to him from a tape recorder or through earphones during a proficiency test. Such tests have face validity, but they need to possess other types of validity as well.

4.3 Reliability. Although I am not discussing the notion of test reliability, or consistency of test performance, in this paper, a brief point must be made about the relationship of reliability to the validity of a test. Whereas it is important that a valid test should possess reliability, the fact that a test has a reliability coefficient of .9 or more does not mean that the test is necessarily valid. We must always stop and first ascertain whether the test is basically doing the job it is intended to do. If it is, we can check its reliability, and seek to improve it if necessary.

5. Application of validity criteria

Carroll (1973) has recently described validity as one of the four persistent problems of foreign language testing. It would appear to be the most important one to tackle and one which will be almost impossible to resolve. It is ignored at both the constructor's and consumer's peril. During the construction phase and during subsequent use validity criteria should be constantly applied to measures, particularly when they have an important bearing on an individual's future. The subjective assessment procedure and the English Proficiency Test Battery ('Davies Test') are two measures used extensively overseas - an ETIC survey recently found that about 5,000 students were assessed by each method in British Council offices in 1973/4 - and validity criteria need to be applied to those measures constantly. This is particularly so in the case of the subjective assessments, since in 90 per cent of the cases the test is unique, ie two administrations of the test, including the oral, are almost never identical. On the other hand, the 'Davies Test' is standard for all candidates, but the predictive power and the concurrent validity of both need to be monitored constantly.

5.1 Subjective assessment. Application of the validity criteria outlined in section 4 above will lead to certain observations and questions. British Council Overseas Standing Instruction 2/72 specified that the four skills should be tested. In terms of construct validity this is satisfactory. Assuming these skills are tested in very general terms we would want to know whether they were being tested separately or in combination. For example, is there a separate text to be read silently

on which questions are set, or is reading part of an oral/aural/reading interview procedure? We would want to know whether the candidate has given sufficient evidence of his ability in each skill for a reasonable assessment to be made. In terms of content validity we would want to know what kind of topics the candidates are reading or writing about and what kind of questions they are being asked to talk about? Is the language relevant to their particular academic or professional and personal pursuits? How much English have they been confronted with during the test and how much has each candidate actually uttered or written? Further practical points are - how much time is needed to assess each candidate, and is there enough time to construct a test soundly? Where large numbers are regularly assessed in this way in a single office this could lead to the emergence of a home-made battery of tests. It might be helpful if testers kept some of these points as a kind of checklist for each candidate. This would also help towards improving the reliability of what is undoubtedly a very unreliable form of evaluation.

5.2 The 'Davies Test' Davies (1965) has presented the evidence for the validity of his test. Chaplen (1970) has argued against the validity of the battery. The TOEFL test consists of five parts - listening, vocabulary, reading comprehension, structure, writing ability - but in a recent follow up three main factors, listening, structure (with writing) and reading comprehension (with vocabulary) were identified. The Davies Test is of a similar vintage and three factors have been identified, segmental listening (test 1), textual listening (test 2) and reading comprehension (tests 3, 4 and 5). Structure is linked with reading comprehension but the subtest formats appear to be more akin to those used in the structure and writing sections of TOEFL. But for both tests we must decide whether they add up to language proficiency or only part of language proficiency. Neither TOEFL nor Davies really provide tests of the productive skills. In fairness it must be said that this has been admitted, and that from the outset Davies has stressed that the battery should be complemented with a writing task and an interview. The language content of the Davies Test is based largely on general English that might be used by a student in both academic and student life situations.

5.3 Both the Davies Test and the subjective assessments could, in certain countries, be supported by other concurrent measures eg performance in local school or university examinations or reports from a candidate's tutor, provided these were thought to be reasonably valid. Face validity for both measures is likely to be linked to their content. Some individual items in, for example, test 4 of the Davies Test may well appear to be unsatisfactory but they were included because of their satisfactory performance during the piloting of the test. Whereas it would be wrong to say that any particular notion of validity is more important than another, it is true to say that in language proficiency testing predictive validity is of particular relevance. Does the test work? Has it predicted satisfactorily? These questions in particular have interested me since both the Davies Test and the subjective assessments have been used for some years. I have been carrying out a survey during the past twenty-four months to try to find the answers to those questions in respect of both measures.

6. Predictive validity criteria

A fairly large sample of overseas students in Britain was needed. Brief details of each individual's general educational background, his or her studies in Britain, and the English language assessments or test scores obtained before leaving their home countries were collated. In addition some criterial information on their ability to use English for their studies while in this country had to be obtained.

6.1 Two sample groups of students have been used, one for the year 1973/4 and one group for 1974/5. All subjects in the sample were first year British Council scholars, Technical Assistance study fellows or Commonwealth Education study fellows who had arrived between approximately June and January and who were studying in Britain for six months or more. The 1973/4 sample numbered 833, of whom 439 had taken the Davies Test and 394 a subjective assessment. The 1974/5 sample numbers 930. Background data and English test scores and grades were obtained from the appropriate departments in the Home Division of the British Council, London.

6.2 Choice of a suitable criterion involved firstly the question of whether to take end of course results or assessments of students' actual language proficiency while studying in Britain. I decided on the latter for three reasons:

a. the difficulty of getting results from university and other authorities and the time this would take following the students' arrival in Britain;

b. the unsatisfactory nature of the information that could be expected. Many students studying on two- or three-year courses might well take no examinations and be authorised to continue on the basis of work done. Others might be on attachment, taking a course, or doing research with no examination in view. Others might have passed only at a resit;

c. I was basically interested in the adequacy of their ability in the use of English and not their academic results. Some students may be unsuccessful for reasons other than English proficiency, and others might well 'pass' in spite of their weak English.

6.2.1 The next alternatives to be considered were whether proficiency should be assessed by tutors' ratings or by another test? If by another test, should it be by an existing one - in which case which test? Or should there be a new test; in which case what would be its format and criteria? I decided on the method of tutors' rating on the grounds that a supervisor usually knows his students quite well and should have had a reasonable experience of their language ability after about six months. I judged that the likelihood of a good response was high if I devised a relatively simple questionnaire involving a minimum of effort on the part of each tutor. I rejected the idea of testing, or attempting to test, the entire sample on grounds of logistics and practicability. However, I did decide to attempt to test a sub-sample of the 1974/5 group as a check on the tutors' ratings. My target was 100 students or approximately 10 per cent of the year's sample.

6.3 Language ability rating. The assumption was made that there is a level of language proficiency adequate for a specific task which is unique to each individual. This level is dependent:

a. on the task; in the case of EAP on the subject of study and on the type of course being taken, and

b. on the individual's general and cultural background and his ability as a language user in that environment, and

c. on the individual's competence in English, his knowledge of the rules of English and of the rules of the community in which he will be using the language.

6.3.1 The tutor was asked to match a. with c. in the case of each student for whom he was responsible and whom I had indicated as being a member of the sample. Basically the tutors were asked to decide in each case whether the student's ability in English was adequate, inadequate or borderline (only just adequate), and, if not borderline, the degree of adequacy or inadequacy. This question was answered first in general terms by placing X in an appropriate box out of six relating to levels of English language ability and then in terms of individual language skills by placing X at an appropriate point along a cline with one extremity representing 'completely adequate' and the other extremity 'totally inadequate'. In addition they had to record whether they thought there had been considerable, some, or no improvement in the student's English. In a questionnaire such as this the problem of reliability is bound to arise, particularly as each tutor is likely to be assessing according to his own criteria. But there were checks built in the questionnaire. Firstly, at least 60 per cent of the tutors did differentiate between abilities in different skills, and, secondly, when their skills ratings were converted to a numerical scale and compared with the general ratings checked in the boxes on a 1 to 6 scale the correlation was around .83. This indicates a high degree of overall consistency in spite of a few maverick individual tutors.

6.4 Communicative proficiency measure. This is the name I have given to my own test of the sub-sample of the 1974/5 group, see 6.2.1 above. I set out to come to conclusions on each student in respect of three questions:

a. Can we establish communication? With what degree of success of effectiveness?

b. Can the student communicate information and ideas to a reader? ie through the written mode.

c. Can the student communicate with written texts? ie do writers communicate information and ideas to the student?

6.4.1 I tackled a. by means of an interview lasting an average of 15 minutes. I sometimes took as little as 10 minutes with a very fluent speaker and 20 minutes or more with a more hesitant speaker, or if time was not pressing and the student was eager to communicate! The interview covered the same topics with each student and the purpose was to elicit information and opinions about the students' work at home, and their life and studies in Britain so far. Care was taken to make the occasion as informal and personal as possible with students giving me information about their studies and background of which I was quite genuinely ignorant.

6.4.2 Question b. was approached by having students write up to 4 or 5 paragraphs on some aspect of their experience in Britain which had

particularly interested them, to describe this experience and give reasons for their interest. Although a very general topic, it was related to the direct experiences of the students. In those cases where they felt there was nothing to write about, alternative topics were agreed, but the principle of expression of information and opinions was adhered to as much as possible. Students were given 30 to 35 minutes for this task.

6.4.3 The third question, communication with written texts, involved the students attempting four cloze passages, each of about 200 words in length, and a total of 100 deletions. Up to 35 minutes were given for this task. Texts covered the topics of agriculture, sex technology, learning mathematics and the sighting of a comet. One passage was chosen from a Ladybird book, one from a mathematics journal and two from the University of Edinburgh Bulletin. It seems to me that cloze procedure, essentially for determining the readability of a text, offered a suitable form of interaction with texts which could be measured objectively. I acted as listener, in both the interview and as reader for the essays and was forced to rely on my own consistency of assessment.

7. Findings of the survey

The findings reported below relate to the 1973/4 sample unless otherwise indicated. The current 1974/5 sample has not yet been processed fully.

7.1 Characteristics of the sample. The basis of selection and breakdown of numbers were given at 6.1 above.

7.1.1 Age. Average age was 30 years.

7.1.2 Countries of origin

Latin America	30% of total
Arabic speaking	19%
E Asia (Nepal, Burma, Thailand etc)	17%
N & W Europe	8 %
Remainder (including E Europe, Ethiopia, Turkey)	26%

Since it is still not the practice to assess English in many Commonwealth countries, few students from Commonwealth countries in East, Central and West Africa are included, and almost none from the Indian sub-continent, Malaysia, Singapore and Hong Kong. There is, however, evidence from other sources that students from these countries are not necessarily all free from serious deficiencies in their ability to use English.

7.1.3 Education. Most are holders of a first degree or professional training qualification at post-secondary level eg teachers.

7.1.4 Subjects being studied in Britain. Subject areas with approximate percentages are as follows:

	1973/4	1974/5
Social Sciences - academic and professional	27½%	26%
Engineering/Technology	17	25
Physical and Biological Sciences	14½	12
Medical	11	7
Agriculture and Veterinary Sciences	8	8
Education and TEFL courses	12	13
English/Applied Linguistics/Arts etc	9½	8

There appears to be a swing this year to the engineering and technological sciences. The other changes, particularly in the general area of medicine, are more likely to be features of the method of sampling rather than a reflection of any significant change away from any particular discipline.

7.1.5 Qualifications aimed at. Approximately two thirds of both sample groups are studying for Masters degrees by instruction or for a postgraduate diploma, certificate or similar non-degree qualification. In 1973/4 22 per cent of the sample were doing research, often in the context of PhD or MPhil studies. This figure has dropped to 13 per cent in the 1974/5 sample. A small increase from 5 per cent to 9 per cent has been noted in the number of sponsored students taking first degrees. About 8 per cent of the sample are on professional or practical attachments.

7.1.6 Remedial English. Of the 1973/4 sample 57 per cent were given pre-session remedial English on arrival in Britain at an average of 8 weeks per student. The proportion has increased to 62 per cent for the 1974/5 sample and the average period of tuition has increased to 8½ weeks. These changes are mainly due to the inclusion in this year's sample of 60 Commonwealth Education Award Fellows, most of whom had lengthy periods of remedial English.

7.1.7 English language assessment overseas. Davies Test scores and subjective assessment grades are set out below. In the case of the subjective assessments the alphabetical grades for the skills were converted to figures, summed and then divided by four to obtain an average grade, and reconverted to an alphabetic grade. The divisions in the tables below are made according to the recommendations for the interpretation of scores on the back page of the British Council's explanatory leaflet on the test battery, and the groupings of the subjective assessments are my own.

Davies Test Score	1973/4	1974/5
less than 34.0	11%	16%
34.0 to 39.9	51	51
40.0 and above	38	33
Mean score	38.9	38.4
Subjective Assessment		
less than B-	13%	12%
B	22	29
B+, A-, A	65	48
Incomplete assessments all levels	-	11
Mean grade	B-	B+

Although there is no significant difference between the mean scores of the 1973/4 and 1974/5 samples, there is a small increase in the number of students with scores of less than 34 on the Davies Test and a decrease in the proportion of students with grades of B+ and above.

7.2 Tutors' ratings. Since the ratings for 1974/5 have not all been processed, only findings for the 1973/4 sample are reported. 85 per cent of the questionnaires were returned, a very high level of cooperation which I much appreciated. Of the students reported on only 4 per cent were deemed by their tutors to have inadequate English for their studies. The dividing line between adequate and inadequate is not simple to draw, of course, and the ratings were re-examined and a new category of 'borderline' was set up on the basis of the tutors' ratings of the language skills as well as on the basis of the 6-point scale of general ratings - see para 6.3 above. The following table emerges:

Adequate	80%
Borderline	16%
Inadequate	4%

About 27 per cent of the sample continued to receive some (concurrent) tuition in English once their main studies had begun. All showed at least some improvement, according to their tutors. But so did almost all the students who did not receive further English tuition and who were not of a very high standard on arrival.

7.3 Communicative proficiency assessment. My own assessment procedures were only carried out on a sub-sample of 96 students, just 10 per cent of the 1974/5 sample. Initial comparison of my assessments with those of the tutors at the levels of generally inadequate or generally adequate shows that our judgments clearly differ in six cases and appear to differ, but not so obviously, in a further six cases. In all but three cases my assessment was the more favourable one. Thus the results of these measures provide evidence to confirm the general conclusions of the tutors' ratings.

7.4 Relationship between predeparture assessments and tutors' ratings

7.4.1 The correlation between performance on the initial proficiency measure and supervisors' judgments as expressed on the skill by skill scale and the general 6 point scales was .32 for the Davies Test and .35 for the subjective assessment. These correlations are low, but they at least indicate some positive relationship between the two measures. These results are in keeping with Sen's findings (1970) when she concluded that English language proficiency has little relevance to final academic performance. Can we expect students to be ranked in approximately the same order by supervisors as they were by the Davies Test or subjective assessment? Furthermore we have to remember that on average a period of ten months has elapsed between taking the test overseas and beginning a course of study in Britain. During this time a lot may have happened, including attending a three-month intensive English course. In view of the fact that the supervisors' ratings were made a further six months after studies had begun in most cases, a correlation of above .3 is perhaps the most that can be expected in these circumstances. Support for this point and for the preceding point on the

Although there is no significant difference between the mean scores of the 1973/4 and 1974/5 samples, there is a small increase in the number of students with scores of less than 34 on the Davies Test and a decrease in the proportion of students with grades of B+ and above.

7.2 Tutors' ratings. Since the ratings for 1974/5 have not all been processed, only findings for the 1973/4 sample are reported. 85 per cent of the questionnaires were returned, a very high level of cooperation which I much appreciated. Of the students reported on only 4 per cent were deemed by their tutors to have inadequate English for their studies. The dividing line between adequate and inadequate is not simple to draw, of course, and the ratings were re-examined and a new category of 'borderline' was set up on the basis of the tutors' ratings of the language skills as well as on the basis of the 6-point scale of general ratings - see para 6.3 above. The following table emerges:

Adequate	80%
Borderline	16%
Inadequate	4%

About 27 per cent of the sample continued to receive some (concurrent) tuition in English once their main studies had begun. All showed at least some improvement, according to their tutors. But so did almost all the students who did not receive further English tuition and who were not of a very high standard on arrival.

7.3 Communicative proficiency assessment. My own assessment procedures were only carried out on a sub-sample of 96 students, just 10 per cent of the 1974/5 sample. Initial comparison of my assessments with those of the tutors at the levels of generally inadequate or generally adequate shows that our judgments clearly differ in six cases and appear to differ, but not so obviously, in a further six cases. In all but three cases my assessment was the more favourable one. Thus the results of these measures provide evidence to confirm the general conclusions of the tutors' ratings.

7.4 Relationship between predeparture assessments and tutors' ratings

7.4.1 The correlation between performance on the initial proficiency measure and supervisors' judgments as expressed on the skill by skill scale and the general 6 point scales was .32 for the Davies Test and .35 for the subjective assessment. These correlations are low, but they at least indicate some positive relationship between the two measures. These results are in keeping with Sen's findings (1970) when she concluded that English language proficiency has little relevance to final academic performance. Can we expect students to be ranked in approximately the same order by supervisors as they were by the Davies Test or subjective assessment? Furthermore we have to remember that on average a period of ten months has elapsed between taking the test overseas and beginning a course of study in Britain. During this time a lot may have happened, including attending a three-month intensive English course. In view of the fact that the supervisors' ratings were made a further six months after studies had begun in most cases, a correlation of above .3 is perhaps the most that can be expected in these circumstances. Support for this point and for the preceding point on the

inappropriateness of the correlation formula is provided by the correlation of just under .3 yielded by the data on students who had had no remedial English, and who had arrived in Britain only a few days before commencing their studies. The most satisfactory correlations, between .5 and .62, were obtained by the following country groups and subject groups assessed by the Davies Test: Mexico, Indonesia, Thailand, Turkey, Peru (subjective as well), Chile (subjective only), medicine, education etc, English language etc, social sciences (academic).

7.4.2 When the pre-departure assessments and supervisors' ratings are examined in another way we get different results. If we group the students into a. those whose English proficiency is considered to be adequate on the basis of pre-departure assessments, with or without remedial English, and b. those whose English is considered inadequate, and then re-group them into x. those whose English proficiency is deemed adequate - including borderline - by their tutors, and y. those whose English is deemed inadequate, we find that 83-85 per cent of the total sample fall into both categories a. and x., 3-3½ per cent into categories a. and y, 10½-13 per cent into categories b. and x., and 1 per cent into categories b. and y. as per the following tables:

Davies Test

Tutors	a	b
x -	85%	10½%
y -	3½%	1%

Subjective

Tutors	a	b
x -	83%	13%
y -	3%	1%

On the basis of these tables, therefore, pre-departure measures have predicted English language proficiency with 84-86 per cent accuracy. In 10½ per cent to 13 per cent of the cases tutors have accepted the English ability of students whereas the pre-departure measures predicted inadequacy. This is satisfactory from the individual student's point of view. What is less satisfactory is that approximately 3 per cent of the students were adjudged to have satisfactory proficiency before they came to Britain, but were subsequently deemed to be inadequate by their tutors. The question now is, how much lower can the cut-offs go before this category of student (a, y in the tables above) increases in proportion? Or is this already too high a proportion?

7.5 Differences according to subject area. The table below summarises the differences, if any, between Davies and subjective assessments combined (in the centre) and tutors' assessments (on the right).

Subject areas	Davies & subjective	Tutors
English language, literature, applied linguistics, linguistics	Well above average	(Well) below average
Education, including TEFL,	Above average	Below average
Arts, Social Sciences (academic)		(Above) average
Arts		Above average
Pure Sciences	Average	Above average
Medicine		Above average
Agriculture	Below average	Average
Engineering, Technology		Below average
Social Sciences (professional)		Below average

The most dramatic difference is registered by the student of English language, literature, linguistics and applied linguistics. Numbering only 52 it is one of the smallest groups, but supervisors clearly demand a very high standard of English proficiency. Students who had averaged a score of 44.4 on the Davies Test were rated as only adequate or barely adequate by their supervisors. Students of Education, including TEFL, are in a similar position. Tutors' ratings are much lower than proficiency test scores would lead us to expect. It is also noteworthy that students of engineering and technology and social sciences (professional), eg urban planning, public administration, arrive with a generally lower level of English which is not altogether satisfactory in the eyes of their supervisors.

7.6 Differences according to type of course. When the sample was analysed according to type of study it was found that students undertaking research arrived with a higher mean assessment and test score and that they received a higher mean rating than other groups. There is every indication that students enrolled in one year Masters and postgraduate Diploma or similar courses have considerable demands made on their English language abilities - heavy lecture load, seminars, long reading lists, frequent essays etc. Furthermore the majority of students of education, engineering, and social sciences (professional) are enrolled in this type of course.

7.7 Remedial English. As expected, there is a difference in Davies Test scores between the group which received remedial English in Britain (mean 36.8) and the group which did not (mean 42.0). The fact is that according to the tutors, everybody's English improved since beginning their studies, except in the case of those who arrived with a very high standard of English. Those who receive remedial English do not appear to catch up with those who have no need of remedial English. Tutors' ratings show that the gap between the proficiency of those who have had pre-session remedial English and those who have not is still very real. There is no doubt, however, that pre-sessional remedial English is of great benefit to those who receive it, particularly those attending special courses put on for university students at eg Manchester, Lancaster, ELTI. One also wonders whether some students may be given too long a period of tuition and whether the content of some of the courses attended is really very suitable.

8. Main conclusions

The above findings have led me to the following conclusions, which in many cases are very much open to further discussion and research.

8.1 The Davies Test, despite weaknesses of construct, is essentially a valid instrument It was developed for students coming to Britain for academic courses. The most satisfactory correlations obtained were by students following the more academic courses, (see 7.4.1 above) ie courses which do not involve laboratory experiments, and manipulation of instruments and machinery, but which concentrate almost entirely on the spoken and written word. Overall it was successful in its predictions of language ability in 86 per cent of the cases and over-optimistic in only only 3 per cent - see 7.4.2 above. There is always the danger of insecure administration, particularly in countries where large numbers are tested.

8.2 The subjective assessment procedures would in general terms appear to be valid - especially when administered rigorously. I sensed that the assessors in Chile knew what was required and were consistent. The Chilean group was the only group of students who were subjectively assessed to obtain a correlation between assessment and tutors' ratings of above .6. Apart from Chile, Brazil and Ethiopia country groups were small in number, there being only four other countries with ten or more students. It is clear that this form of assessment is a more unreliable form and that the grades awarded bear little relationship to their official definitions. Clearly some countries eg Brazil, are consistently lenient. The grading of oral/aural skills is frequently too generous and assessors are often just too familiar with the local accent or variety of English. It can come as a bitter shock to a student once in Britain to find that British people, including his supervisor, are not so ready to adjust to his form of English and that he is not so easily understood.

8.3 Certain disciplines, particularly when studied in the context of a one year's Masters, Diploma, Certificate or similar course, appear to demand a high standard of English ability. These include English language and literature, applied and theoretical linguistics, education and TEFL. There is also an indication that students of the Engineering and Technological Sciences, together with those taking professional courses in the Social Sciences, are thought to be able to succeed with a lower standard of English. Their supervisors' reactions show that this is a dangerous assumption.

8.4 Tutors who know their students well can give a valid judgment on their students' English. Although they may confess to leniency in some cases their comments indicate they are able to separate language problems content or background problems. I am also satisfied that the simple questionnaire devised is a valid instrument.

8.5 With a score of around 37 (or assessment of B), plus remedial English where necessary, a student is likely to have adequate English for his studies. Thereafter, as his score decreases the likelihood of his being handicapped through inadequate proficiency increases, particularly in the disciplines mentioned in 8.3.

8.6 Scores should be interpreted in the context of each individual's total background and his study needs in Britain.

8.7 There is a certain threshold or minimum level of English proficiency which is seen as adequate by the individual and by the tutor, although they may not agree as to when it has been reached! Some improvement seems inevitable once a student is in Britain. Most students will reach their particular threshold and go on beyond. Some will have reached it before leaving their home countries but still make efforts to improve. A number of students will spend a lot of their energy attaining that level and then concentrate solely on their studies. A small number, sadly will find that they cannot reach that minimum level after all, give up trying to attain it, and struggle on somehow gleaning what they can from their courses. The testing and follow-up measures described should yield information which can help us determine how near each individual is to that threshold.

8.8 The testing procedures used, particularly the Davies Test, are essentially 'micro' tests, examining a student's command of the phonology, the lexicon and the grammar of English. The evidence suggests that they provide satisfactory, though by no means optimum, information necessary for taking informed decisions. I have no evidence to suggest either tests geared to ESP or 'macro' tests would do the same with any lesser or greater efficiency. There is plainly room for much more development and research into all these types of proficiency test.

8.9 The findings reported here provide evidence, though not the only evidence, to conclude that the Davies' English Proficiency Test Battery possesses overall validity as a test of English language proficiency. And this in spite of some sections, eg 1 and 2, which have low construct validity and low reliability, and in spite of other sections which have items whose content is questionable. It is possible that the construct of a test such as EPTB would not be considered valid by the specialists if similar specifications were submitted to them today. Yet analysis of some of its performance shows shows it to be overall a valid measure. This finding leads me to propose that a carefully constructed and pre-tested proficiency measure is likely to possess overall validity even if some of its parts exhibit deficiencies.

REFERENCES

CARROLL, J B — 'Fundamental considerations in testing for English language proficiency of foreign students' in Testing the English proficiency of Foreign Students, Center for Applied Linguistics, Washington DC (1961)

" " — 'Foreign Language Testing: Will the Persistent Problems Persist?' in Testing in Second Language Teaching: New Dimensions, Dublin, ATESOL (1973)

CHAPLEN, E F — Identification of non-native speakers of English likely to underachieve in university courses through inadequate command of the language, unpublished PhD thesis, Manchester (1970)

CLARK, John L D — Foreign Language Testing: Theory and Practice Philadelphia, The Center for Curriculum Development (1972)

DAVIES, A — Proficiency in English as a Second Language unpublished PhD thesis, Birmingham (1965)

" " — (Ed) Language Testing Symposium, London, OUP (1968)

PILLINER, A E G — 'Subjective and objective testing' in Davies (ed) (1968)

SEN, A — Problems of Overseas Students and Nurses, Slough NFER (1970)

TOEFL — Test of English as a Foreign Language, Princeton Educational Testing Services.